ON THE ORIGIN OF COMETS
Asoka Mendis and Hannes Alfvén

1. Introduction

The cosmogony of the planetary and satellite systems consists of understanding the physio-chemical processes leading to their formation and also trying to decide at what time and over what period their formation took place. The cosmogony of the comets require answers to not only these two questions but also as to where, in relation to the solar system, the observed and inferred distributions of comets were formed.

One also recognizes that unlike in the case of the larger bodies the time scales of dynamical and physical evolution of some of these bodies are very much smaller than the age of the solar system. This leads directly to the question of the maintenance of their observed abundances and consequently to the genetic inter-relationships between the various classes of comets and also to those between comets and other bodies in the solar system. It also provokes the question whether the formation of the comets was completed long ago together with the rest of the solar system or whether the process of formation may be still continuing even though on a much diminished scale.

Attempts at answering each of these questions has produced a number of interesting ideas, but despite considerable effort by a number of authors it must be admitted that all of these questions still remain
largely unresolved, although the continuing work on the dynamical evolution of cometary orbits have put important new constraints on the evolutionary path of these bodies.

So far various theories have proposed solar origins, protoplanetary origins, planetary origins and interstellar origins. They have also proposed completed past origins as well as continuing origins. Comprehensive reviews of these ideas are available elsewhere \(^{(2),(3),(4),(5)}\). Here we will restrict ourselves to offering a few comments pertinent to some of these problems.

2. Observed and Inferred Distributions

Up to the present time about 100 individual short period \((P \lesssim 200\ \text{yrs})\) and over five times as many long period comets have been discovered, and the present rate of discovery averages about 4 long period and 1 short period comet per year \(^{(5)}\).

The differences in the orbital characteristics between these two classes are well known. The short period comets which spend almost all their time within the confines of the planetary system have mostly low inclination \((i \lesssim 25^\circ)\) orbits. Only five of them are known to be retrograde. Also about 2/3 of them have aphelia close to Jupiter's orbit and are likely to be strongly influenced by that
planet. The long period comets on the other hand show a uniform distribution in inclination with about equal numbers having prograde and retrograde orbits. They are also for the most part moving in almost parabolic elliptic orbits with periods in excess of $10^6$ yrs. Based on a statistical analyses of 22 long period comets whose original barycentric orbits had been accurately calculated, Oort showed that the bulk of them seemed to come from a region between about $3 \times 10^4$ A.U. and $10^5$ A.U. with a median value of about $5 \times 10^4$ A.U. He also noted that average planetary perturbation in $1/a \left( < \Delta \left( \frac{1}{a} \right) > \right)$ which amounted to about $± 5 \times 10^{-4}$ A.U. was more than an order of magnitude larger than the observed dispersion in $1/a$ near the maximum. He was thus led to conclude that the observed long period comets were "new" in the sense that they were being observed at their first passage through the inner regions of the solar system ($q \lesssim 2$ A.U.). Based on the frequency of discovery of new comets, their average period and an assumed distribution of the transverse velocity at aphelion Oort further deduced that the number of "intrinsically observable" comets in this reservoir ($3 \times 10^4 \leq Q \leq 10^5$ A.U.) must be in excess of about $10^{11}$. Although Oort's conclusions have been strongly criticized by Lyttleton, a more recent detailed analysis by Marsden and Sekanina seems to confirm them,
despite the very small numbers on which the statistics are based. They have shown that for comets having perihelion distance more than 3 A. U. and which are thus likely to be free of non-gravitational forces if their volatile component is largely water ice as is now generally believed\(^9\), the distribution of original barycentric orbits show a remarkable concentration corresponding to an aphelion distance around \(5 \times 10^4\) A. U. Of course if these "new" comets are charged with a component much more volatile than water or the clathrate\(^{10}\) then this result too could be largely fortuitous.

Besides these distributions one has to grant the possible existence of others. Indeed a comet having aphelion \(\leq 2 \times 10^4\) A. U. and perihelion well outside the planetary system will be dynamically stable against both stellar and planetary perturbations, during the lifetime of the solar system. It may also be barely possible to have some comets stably trapped in certain peculiar orbits in the outer regions of the planetary system over the cosmogenic time scale\(^{11}\). Furthermore it is known that there is a continuous ejection of long period comets from the solar system at the present time and the process may have proceeded on a grander scale during the formation stages of the solar system. Consequently interstellar space may be continuously being populated by comets from our own solar system as well as others like our own. We shall, however, concern ourselves here mainly with the observed distributions.
3. The Origin of Long Period Comets

Believing that it was difficult to form comets in situ at such large distances ($r \approx 5 \times 10^4$ A.U.) Oort suggested that they originate within the inner solar system. They were ejected out by planetary perturbations and while some would have immediately escaped the solar system in hyperbolic orbits those on elliptic orbits whose aphelia $Q, > 10^5$ A.U. were subsequently removed by stellar perturbations over a time scale of $5 \times 10^9$ yrs, whereas those with $Q \leq 2 \times 10^4$ A.U. are hardly affected at all. These two values of $Q$ define the limits of the so-called Oort's cometary reservoir. Oort further showed that while stellar perturbations will completely isotropize the velocity distribution near aphelion of comets in this region, the continual reshuffling of the velocity distribution will continuously inject some long period comets into orbits bringing them to the vicinity of the sun to explain the observed isotropic distribution. Although Oort originally made the highly unlikely supposition that these comets, together with the minor planets resulted from the break up of a planet inside Jupiter's orbit, several other authors have subsequently suggested that these comets originate in the outer regions of the planetary system in a more reasonable way. (12), (13)
The difficulty with this scheme is already apparent from Everhart's calculations (14) for the diffusion of the $1/a$ values of hypothetical comets started within the solar system (despite their incompleteness, particularly the neglect of stellar perturbations). If we, however, accept Everhart's linear law for a number of orbits vs $1/a$ and scale it for the fact that there are, say, $10^{11}$ comets in the region $2 \times 10^4 \text{ AU} - 10^5 \text{ AU}$, this seems to require an embarrassingly large number of comets within the solar system at some time ($\gtrsim 10^{16}$).

Recently Alfvén and Arrhenius (15) have developed a detailed hydro-magnetic, planetesimal theory for the formation of planetary systems around a central star as well as the formation of satellite systems around a central planet. The basic steps in the process are the following: Initially gas infalling towards a spinning, magnetized central body is ionized and brought into partial corotation. Grains condensing out of this plasma fall on neutralization towards the equatorial plane and are collected there at various discrete distances from the central body due to mutual inelastic collisions to form streams of almost co-orbital particles called "jet streams". These grains then further accrete within these streams due to mutual inelastic collisions growing into larger and larger planetesimals which ultimately grow into planets and satellites, the final stages of the accretion process being gravitational.
The central problem here is the time evolution of these jet streams which have been studied recently both numerically (16) and analytically (17) with the authors drawing basically similar conclusions. In order to make the problem tractable a number of simplifying assumptions have been made in both cases. In particular, the effects of fragmentation and accretion have been neglected as are gravitational perturbations and electromagnetic effects such as the Poynting-Robertson effect. Within these limitations, however, one finds in a general way that, if collisions are sufficiently inelastic, a radial focussing or clustering would occur such that the thickness of the stream is reduced.

More recently Ip and Mendis (18) have studied the time evolution of such streams using simple mathematical models which also take into account the effects of fragmentation and accretion. Accretion here meaning not merely the coagulation effect of stream particles sticking to each other during inelastic collisions but also the continuous sweeping up of matter intersecting the streams. The treatment is in terms of the average kinetic and physical parameters of the particles and considers for simplicity a pure accretion case and a special fragmentation case wherein despite the competing effect of accretion, fragmentation continues to keep the average grain radius constant. The results of the computation are shown in the following figures. Figure 1 depicts the pure accretion model. Here \( \Lambda_0 \) is the initial value of the ratio of the accretion time
Figure 1: The variations of the normalized internal velocity, the number density and the grain radius with time, for different values of $\Lambda_0$, in the pure accretion model.
scale to the internal collision time scale, and time is measured in units of
the initial accretion time scale. For $\Lambda_o = 1$ we have a gradual dispersion
of the matter stream due to the thermalization effect of the accretion of the
external matter. In the case of $\Lambda_o = 100$ there is a rapid focussing
of the stream because the evolution of the stream is dominated by the
inelastic collision process among the stream particles. An intermediate
behavior is observed when $\Lambda_o = 10$, the matter stream has an initial
expansion phase until $\tau \sim 1$. At this stage the thermalization effect is
balanced by the internal energy dissipation by inelastic collisions, and
contraction begins. It seems therefore that, in the case of a pure accretion
model for interplanetary matter streams, focussing will always occur if
$\Lambda_o \gtrsim 10$.

Figure 1 also shows that for $\Lambda_o \lesssim 10$ the particle density of the
stream is reduced by three orders of magnitude within a period of about $3 \tau_{A0}$
while the average radius of a grain increases by one to two orders of
magnitude due to the efficiency of the coagulation process in the stream.

Figure 2 depicts the fragmentation model. The variations of $\vec{v}_i$
and therefore the thickness of the stream are similar to those of the pure
accretion model. However, in the case of $\Lambda_o = 10$, the contraction of
the stream occurs much faster culminating in a catastrophic collapse just
before $\tau \approx 2$. Due to the rapid focussing the particle density begins to in-
crease almost instantaneously following an initial phase of gradual decrease.
Figure 2: The variation of the normalized internal velocity and number density with time for different values of $\Lambda_0$ in the fragmentation model.
In the evolution of any proto-planetary matter stream while there would be a gradual increase in the average grain size as shown by the pure accretion model, this growth would be hindered to some extent by the competing effects of fragmentation. Consequently the real situation would be intermediate to those suggested by the two models we have discussed. The general conclusion then is that any proto-planetary matter stream in which accretion and fragmentation are taking place a strong focussing would occur over a period of a few accretion time scales.

While planets and satellites will be formed in this way close to the equitorial plane of the central body, dust particles associated with the gas and having a sufficiently small charge to mass ratio not to be significantly effected by the magnetic field will fall in streams towards the sun. If we consider a spherical cluster of such dust of cometary mass \( \approx 10^{18} \text{ g} \) initially at a large heliocentric distance \( r_A \) from the sun falling in towards it in a highly elongated elliptic orbit, then if \( r_B \) is the heliocentric distance sufficiently before perihelion such that a linear approximation may be made to the portion of the orbit between \( A \) and \( B \), it is seen that the cluster will be drawn out into a thin pencil shaped stream near \( B \) whose length \( \approx \sqrt{\frac{r_A}{r_B}} D_A \) and whose cross-sectional diameter is \( \left( \frac{r_B}{r_A} \right)^2 D_A \). Consequently the density will be increased by a factor \( \left( \frac{r_A}{r_B} \right)^{3/2} \). If we take \( r_A \approx 5 \times 10^4 \text{ AU} \) and \( r_B \approx 5 \text{ AU} \), and the distributed dust
density at $A \approx 2 \times 10^{-20}$ gm cm\(^{-3}\) (corresponding to a neutral gas density $\approx 10^4$ cm\(^{-3}\)), $\rho_B \approx 10^6 \rho_A \approx 2 \times 10^{-14}$ gm cm\(^{-3}\). Since the internal collision time scale at $B$ is given by $t_B \approx \frac{r_g \rho_g}{\rho_B v_{rel}}$, taking $r_g \approx 10 \mu$, $\rho_g \approx 0.5$ gm cm\(^{-3}\), $v_{rel} \approx 10^4$ cm sec\(^{-1}\), we get $t_B \approx 2 \times 10^6$ secs $\approx 1$ mo. Consequently a fast focussing into consolidated body of cometary size is possible during a single perihelion passage. While the isotropy of the observed distribution of long period comets is a natural consequence of this formation process, the emerging view of a comet as a lossely consolidated grainy matrix is consistent with such a formation. It also anticipates the observed compositional similarities between interstellar dust and comets.

It should be noticed that the mechanism we are proposing is essentially different from Lyttleton's gravitational lensing\(^{(19)}\). It is also asserted that these dust streams are unstable against the effects of internal inelastic collisions and would quickly agglomerate into one or more larger bodies.

4. The Origin of Short Period Comets

The idea that short period comets derive from long period ones that pass near one of the massive outer planets (especially Jupiter) and lose energy is nearly two centuries old being generally attributed to Laplace. This classical capture hypothesis has since been considered by several authors and worked out in detail by Newton\(^{(20)}\) whose calculations have
been extended and refined more recently by Everhart\textsuperscript{(21)}. Both authors reached the conclusion that single close encounters of long period (or more precisely parabolic) comets belonging to the observed random distribution, with planets (particularly Jupiter) cannot solve the problem of the origin of short period comets. While the capture probability remains finite although very small, the calculated post-capture distribution of these short period comets following a single close encounter with Jupiter does not in any way correspond to the observed distribution and nowhere is this discrepancy more marked than in their distribution with regard to period and inclination. In fact, these calculations predict that about a quarter of the short period comets with perihelion \( \leq 2 \) A.U. and period \( \leq 21 \) yrs should have retrograde orbits although there are none observed.

Very recently Everhart\textsuperscript{(22)} has made a Monte Carlo statistical study of the interaction of hypothetical random parabolic comets with the Sun-Jupiter system, following some comets up to 2000 returns. While elucidating several important points regarding the capture hypothesis it identified a so-called "capture-region" consisting of prograde comets of low inclination (\( i_0 \leq 9^\circ \)) having perihelia close to Jupiter's orbit (\( 4 < q_0 < 6 \) A.U.) from which over 90\% of the captures take place. While the calculated post-capture distributions agree rather well with observation, subsequent work by the same author\textsuperscript{(14)} shows that from a purely orbital evolution point of view the source of the observed short period distribution could
equally well be situated within the confines of the solar system, in particular the Jupiter-Saturn region. Consequently such calculations have so far not succeeded in unambiguously identifying the source region of the observed short period comets. The problem of maintaining the observed short-period population against dissipation and fading require a capture rate of at least one every ten years or so. Based on the rate of capture deduced from Everhart's numerical results and the deduced rate of injection of "new" comets from the Oort cloud into the "capture region" Joss concluded that the capture rate was 4-5 orders of magnitude too small to account for the observed number of short period comets. Delsemme on the other hand, considering also the intermediate period distribution and assuming a concentration towards small inclination in the capture region concludes, on the basis of the number of comets reaching perihelia per unit time, that no such discrepancy exists. Besides a number of questionable assumptions and the uncertainties in the several parameters inherent in Delsemme's analysis it needs to be realized that the $10^8$ "intermediate period comets" required to increase the capture rate must ultimately derive from the "new" comets entering the capture region from the Oort cloud $(2 \times 10^4 \text{ AU} \leq \mathcal{Q} \leq 10^5 \text{ AU})$ due to stellar perturbations. Even with an orbital diffusion time scale as large as the age of the solar system one still needs an input of such comets at the rate of 1 every 50 years. Consequently we now have a problem not of accounting for the observed short period comets
but rather for the deduced abundance of "intermediate period comets" from which the short period comets are supposedly derived.

Vsekhsviatsky had earlier attempted to circumvent this difficulty by reviving the old Lagrangean idea of an eruptive origin for short period comets. In a series of papers (e.g. see ref. 25), he has successively proposed that comets are the ejecta of violent volcanic eruptions on the surfaces of planets (particularly Jupiter) and their satellites. Besides the essentially circumstantial nature of the evidence, simple considerations based on the energy requirements as well as the survival of these objects during such violent eruptions argues strongly against such a view.

This leads us to our final topic: the genetic relationship between comets and other small bodies in the solar system. The orbital associations of comets and meteor streams on the one hand and the formal similarity of the orbits of short-period comets and Apollo-type minor planets on the other, have been known for a considerable time. Both classes of objects are generally believed relics of comets. While the Apollo-type minor planets are believed to result from a complete degassing of cometary nuclei and the consequent shrinkage of their orbits due to non-gravitational forces, meteor streams are believed to result from the complete or partial disintegration of cometary nuclei.
There are at least 17 major permanent meteor streams observed to intersect the Earth's orbit, while a somewhat smaller number of temporary meteor streams too have been observed. Since meteor streams (whose typical thicknesses are $\lesssim 0.1$ A.U) are observed only when their orbits are favorably positioned with respect to the earth's orbit, the total number within the solar system is likely to be much larger. Several of these meteor streams are known to be approximately co-orbital with comets (e.g. Perseids with P/Swift-Tuttle, October Draconids with P/Giacobini-Zinner, Leonids with P/Temple-Tuttle, Taurids with P/Encke, etc.). It is of course not unreasonable that meteor streams should be considered as the disintegration products of comets since we witness cometary erosion--i.e. the loss of gas (type I tails) and dust (type II tails) as comets approach the Sun, all the time. We have also seen on several occasions cometary disintegration; P/Biela was seen to break up into two parts in 1846. Subsequent to break up both comets moved in very close orbits and were seen at their next return in 1852 separated by about $3 \times 10^6$ km. They were never observed after that, but a temporary meteor stream (Andromedids) is now believed to be associated with their orbit. While we cannot deny a process occurring before our eyes, we need not necessarily assume that this process is irreversible. It seems worthwhile considering, in the light of the development of the theory of jet streams, whether the opposite process, viz., comets and/or Apollo-type minor planets forming in meteor streams is also possible.
A meteor stream, where we have a swarm of particles moving in Kepler orbits in a gravitational field with a small spread both in velocity and configuration space is a good physical example of an idealized "jet" stream suggested by Alfvén and Arrhenius, and whose dynamical evolution has been discussed in section 3. Trulsen\(^{(26)}\) has made a preliminary study of the effects of planetary perturbations on meteor streams. He considers the case of Jupiter producing a perturbation in a co-orbital eccentric jet stream of particles dispersed along the orbit. A velocity modulation is produced which causes a traveling density wave. The longitudinal focussing achieved this way is, of course, only temporary for any given group of particles and the maximum focussing achieved is only about 20. A greater compression could possibly be obtained through the interference of two such waves excited at consecutive close approaches of the meteor stream to the planet. Besides, if viscous effects of some form are present, as would be the case if an appreciable quantity of gas could be retained in the stream for a sufficient time, it may be possible to achieve a more permanent condensation which may be considered as the birth of a comet or at least an Apollo-type minor planet. The process considered above must be of a rather frequent occurrence because only a modest modulation is required to trigger it. In fact, if the modulation is too large, as would be the case if the stream approached Jupiter too closely, it is a scattering rather than a focussing that results.
Mendis(27) has considered the time scales for a number of dispersive effects including differential precession of nodes and perihelia, the dispersion of particles of different sizes due to the Poynting-Robertson effect, the longitudinal dispersion due to variation of the "effective" gravity on particles of different sizes moving in the combined gravitational and radiation fields of the Sun, and the dispersion due to the differential efficiencies of accretion of particles of different sizes. It is found, in a typical case, that all these times are comparable or larger than the time scale for agglomeration, which is typically about $9 \times 10^4$ yrs with the typical values adopted by Ip and Mendis$^{(18)}$: $V_{10} \approx 0.5$ km/s, $r_g \approx 10\mu$, $a \approx 3$ A.U., $\rho_i \approx 10^{-22}$ g/cm$^3$, $\rho_s \approx 10^{-20}$ g/cm$^3$, $\lambda_0 \approx 10$.

The low elasticity and high sticking coefficients assumed in these calculations seem to be supported by the studies of the surface properties of lunar dust grains, dust grains sticking to the protective paint of Surveyor III and also of dust grains artificially irradiated with large doses of low energy particles simulating solar wind conditions$^{(28)}$. Furthermore if a factor is allowed for the clumpiness of these streams the focusing time scale would be further reduced, so will the retention of a sufficient quantity of gas. Consequently while the situation remains somewhat marginal an eventual focusing of some present day meteor streams is not excluded. The initial expansion phase noticed in our recent model computation (see figures 1 and 2) too is interesting in that it may explain the claim that "young" meteor streams are dispersing faster than can be explained by planetary perturbations or electromagnetic effects$^{(29)}$. 

655
It is also shown\(^{(27)}\) that unlike comets, meteor streams could be very efficient in accreting matter from interplanetary space due to their large "effective" cross-sections. How much could be collected of course naturally depends on the highly uncertain dust density of the interplanetary space, especially in the regions beyond Jupiter. However, should this be even as much as two orders of magnitude lower than the distributed density in meteor streams a fraction of about \(10^{-5}\) can be collected by the stream per revolution, which could account for the volatile fraction in the subsequently consolidated comet, if a significant fraction of these interplanetary grains contain such a component, perhaps in the form of clathrates.

An interesting observation in this connection concerns P/Temple-Tuttle \((P \approx 33.2\) yr\) which was first recorded as a diffuse but bright object only as recently as 1866\(^{(30)}\) although the associated Leonids had been known for centuries earlier. An even more significant observation concerns Comet \(P/{\text{Swift-Tuttle}}\) \((P \approx 120\) yr\) which was bright on its first apparition \((in\ 1862)\) to be easily seen with the naked eye being a 2nd magnitude object at its brightest\(^{(27)}\). What is surprising is its association with the Perseids meteor stream which has been observed for over twelve centuries\(^{(31)}\). Both these observations seem to indicate that comets may have formed in already existing meteor streams. Due to the very large times which span these observations and the uncertainty with regard to the conditions of the early observations we hesitate to draw any strong conclusions from them at this stage except to state that they seem very suggestive. It should, however, be stressed that these observations are of such an important nature that their significance merit further investigation.
If indeed the genetic relationship between comets and meteor streams is a reciprocal one with meteor streams providing not merely a sink for comets but also a source, it could very well mitigate the crucial difficulty at the present time, with regard to the observed abundance of short period comets. At a more basic level is the intriguing possibility that the comet-meteor stream complex may provide us with a cosmic laboratory where we could still observe even though on a much diminished scale the planetesimal process which led to the formation of the solar system over 4.5 million years ago.
References

(1) Everhart, E., 1974, paper read at the 25th IAU Colloquium: The Study of Comets, GCFC, Greenbelt, Maryland (Oct.-Nov. 1974).


(15) Alfven, H. and Arrhenius, G., 1970, Astrophysics and Space Sci., 8, 338 (paper I); 1970, 9, 3; (paper II); 1973, 21, 117; (paper III); 1974, 29, 63 (paper IV).


